Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/46/20241517

The prediction of suicide rate based on machine learning

Tianzhuo Luo

Basis International School, Parklane harbor, Songtao Road, Park Lane Harbour, Daya Bay, Huizhou, China

Tianzhuo.Luo14138-biph@basischina.com

Abstract. Suicide rates have a profound impact on individuals and society, making it a significant social issue warranting extensive research. To predict and understand suicide rates, researchers employ various methods, including social science research, psychological surveys, and data analysis. The principal aim of this investigation was to develop a reliable forecasting model that can accurately predict suicide rates and provide credible references for relevant populations. In this research, multiple machine learning and deep learning methods, Neural networks, Random Forests, and Support Vector Machines (SVM) were among the learning techniques used., and XGBoost. These methods were chosen due to their effectiveness in handling complex and multidimensional data. The results demonstrated that a comprehensive analysis of multiple indicators was crucial, with neural networks yielding the best performance and SVM performing the worst. These research findings offer valuable insights into the factors influencing suicide rates and can inform the formulation of prevention and intervention strategies.

Keywords: Machine learning, suicide rates, Xgboost, SVM, Random forest, Neural Network model.

1. Introduction

In the recent decades, the problem of suicides gradually becomes a severe issue, across the world. Different factors can serve as causes towards suicide, due the rapid development of technologies, people started to encounter abundant pressure, if they do not work and study diligently, they will be eliminated by the society. However, the suicide rates vary throughout the world. Some countries suffered from high suicide rate and some countries suffered less. Pressures are not the only factor of committing suicide, other factors, such as emotional status, financial status, psychological status, and environmental conditions can as well lead to the occurrence of committing suicide.

In order to conduct the research and prediction of the suicide rates, scientist conducted variety of modeling method to acquire the result, such as multivariable logistic regression, Box-Jenkins Model by utilizing the model of autoregressive integrated moving average from previous years in order to predict the suicide rate in the upcoming years, and the application of XGboost and SVM. Current study of suicide has shown that in all suicide's occurrence, over 77% of suicide happens in the low- and middle-income countries. The major cause of suicide is mental disorders, such as Schizophrenia, bipolar disorder…etc. Another crucial factor of suicide occurrence is discrimination, like the victim of racial discrimination, part of the LGBTQ group...etc. Despite the abundant of research has been done on suicide rates, but however, certain problems like data qualities might affect the result of the data. Due

© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/46/20241517

to the lack of data collection techniques in certain low income countries, or due to the differences on the method of data collection. So the current researches has just shown a trends of what the real suicide rates are alike.

The study and investigation of suicide rate are crucial for the development of the society. By using machine learnings like the random forest, XGBoost, SVM, Neural Networks, these methodologies helped with more precise prediction of the suicide rate, which can lead to the study of a variety of aspects of committing suicide. By the more precise predictions, the occurrence and patterns of suicide can be determined, by the study of those patterns of occurrence of suicide, the causation of the suicide could also be identified. Identifying the patterns and the causation of suicide were very beneficial on intervene the happening of future suicides been committed. According to statistics, 1.7% of death happens because of suicide, that's over 700,000 people, by aiming for the decrease of suicides, the society will benefit from steady population rates, globally.

In order to provide a comprehensive comparison analysis, this study selected four different methods for examining the prediction of suicide rates: Random Forest, XGBoost, SVM, and a three-layer Neural Network model.

2. Method

Different methods are used in order to conduct the research of this topics. Such as Random Forest, XGboost, SVM, Neural networks. For each of them, they have different functions, but all aimed to acquire a precise predictions of suicide rates. SVM is an algorithm to determine the N-dimensional space that categorizes different datasets, to locate the plane that has the maximum difference. Xgboost is used to deal with such large dataset and has the ability to deal with missing values. By ensemble multiple decision trees, random forest method can generate are prediction with higher accuracy. By using the Neural Networks which mimics the function of the animal's brain, it can self-learn and generate more appropriate decisions.

2.1. SVM

The goals of application of the SVM are an algorithm to determine the hyperplane in the N-dimensional space that categorizes different datasets. One crucial objective of creating SVM is to locate the plane that has the maximum difference [1] (The biggest discrepancy between the datasets with the highest and lowest values). Hyperplane, is the decisive boundaries that classify datapoints. The dimensions of the hyperplane are relied on the amount of features. Support Vectors, they are the datapoints that's near to the hyperplane, and effects the location and the initial point of the plane. By applying the hinge loss, the margin can be maximized [2].

$$min_w \|w\|^2 + \sum_{i=1}^{n}(1 - y_i\langle x_i, w\rangle)+ \tag{1}$$

By applying regularization parameter, it helps to balance the maximization of the margin and the loss. So it looks like

$$c\big(x, y, f(x)\big) = \begin{cases} 0, if\ y * f(x) > 1 \\ 1 - y * f(x), else \end{cases} \tag{2}$$

2.2. XGboost

XGboost, the extreme gradient boosting. It has strong ability to deal with large datasets, and capable of handling with missing values (Figure 1). It's features of parallel processing, making it to train models with large datasets more efficiently. The algorithm that Xgboost use is the gradient boosting decision tree algorithm, through the application of stochastic gradient boosting, multiple additive trees, and gradient boosting [3]. By saying boosting, it means to use an ensemble technique that adds new models to fix the errors the previous ones had made. In more advance, the gradient boosting technique, is to add new models in order to predict the residuals of the previous model and add them together, ultimately to complete the final prediction. The trees are also important to complete the XGboost. Decision trees aims to by applying the rules of decisions from data features, to predict the values of certain variable. Bagging

DOI: 10.54254/2755-2721/46/20241517

classifiers are used to sum up individual predictions and to perform a final prediction. Due to the high variance of the decision tree, combination of different decision trees in a paralleled form, will leads to the resultant variance in to a lower form. The math for XGboost is [4]:

$$G_j = \sum_{i \in I_j} g_i \tag{3}$$

$$H_j = \sum_{i \in I_j} h_i \tag{4}$$

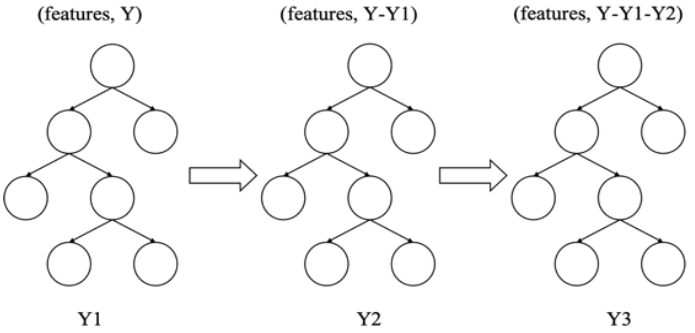

Figure 1. Boosting model

2.3. Random Forest

Random Forest: it is a supervised learning algorithm by ensemble multiple decision trees, using the method of bagging in order to generate more accurate predictions. Random Forest can be used in both classification and regression [5]. The random forest increases the model's randomness in proportion to the number of trees in the model. When splitting a node, only a random subset of the features is taken into account., which increases the diversity and randomness of the trees [5]. The random forest generates the outcomes based on the prediction of the decision tree, therefore, a model's prediction would be more accurate the more trees it included [6]. It can manage missing values in addition to handling numerical and categorical data. It is a potent instrument in the toolset of data scientists and is used in numerous machine learning packages [7].

2.4. Neural Network

Neural Networks are part of deep learning algorithm, they mimics the neural systems of human,neural networks are composed of node layers, input layer, hidden layers and output layers [8]. The nodes serves like neurons in human brain. The nodes only activates when the output of individual nodes is higher than a specific threshold, once activated, they will be able to send data to the next layer [8]. In summary, the data enters neural network at the input layer, which later in the process, they will connect with the hidden layers [9]. The nodes in the hidden layers will combine the data and send to the output layers. The neural networks are adaptive, they manipulate themselves when they learns [10]. Neural networks possess the capacity to identify correlations in data and uncover hidden patterns, throughout processes, neural networks are able to improve themselves. Neural networks can be beneficial, it can learn complicated non-linear models, it can improve the process of decisions on many aspects.

3. Experimental setup and results

3.1. Dataset Overview

A research data published on Kaggle presented the number of suicides committed in different counties, continents by their ages (Table 1).

DOI: 10.54254/2755-2721/46/20241517

Table 1. Data set attributes

Region Name	The continent of the countries of the data
Country name	The nation in which the information is gathered
Year	The year when the data is collected
Sex	The gender of the data sample
X years	The ages of the suicide occurs
No. Suicide	The amount of suicides occurs in the data
Percentage	The percentage of suicide occurs in that population of data

3.2. Evaluation Metrics

The metrics used for model evaluation are MAPE, RMSE, RMSLE and R^2 (adjusted).

- *Mean Absolute Percentage Error (MAPE)*

$$\text{MAPE}=\frac{1}{N}\sum_{t=1}^{N}\left|\frac{A_t-F_t}{A_t}\right| \tag{5}$$

where A_t is the actual value, F_t is the predicted value and N is the number of predictions.

MAPE presents the error in the form of percentages, the smaller the MAPE the better the predictions.

- *Root Mean Squared Error (RMSE)*

$$\text{RMSE}=\sqrt{\text{MSE}}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2} \tag{6}$$

where y_i refer to the actual value, $\hat{y}_i$ represent the predicted value, and N denote the number of predictions.

RMSE represents the standard deviation of the prediction errors (residuals), which evaluates the degree of scatter of these residuals. In other words, it indicates how centralized the data is around the line of best fit. Notice that RMSE value is scale-dependent, which increases significantly if the scale of error increases.

- Root Mean Squared Log Error (RMSLE)

$$\text{RMSLE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}[log(y_i+1)-log(\hat{y}_i+1)]^2} \tag{7}$$

Where it comes to the numbers of predictions, y_i is the actual value, and $\hat{y}_i$ is the predicted value and N is the number of predictions.

As a variant of RMSE, RMSLE transforms the predicted and actual values into a logarithmic scale. Not only can RMSLE be It is resilient to the impact of outliers and is defined as the correlated error between the real and expected values.

-Adjusted R^2

The determination coefficient, or R^2 is defined by

$$R^2=\frac{\sum_{i=1}^{N}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{N}(y_i-\bar{y})^2}=1-\frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{N}(y_i-\bar{y})^2} \tag{8}$$

where $\sum_{i=1}^{N}(y_i-\bar{y})^2$, $\sum_{i=1}^{N}(\hat{y}_i-\bar{y})^2$, $\sum_{i=1}^{N}(y_i-\hat{y}_i)^2$ denote the total sum of squares (TSS), the explained sum of squares (ESS) and residual sum of squares (RSS) respectively.

The adjusted R^2 is calculated by

$$R^2_{adj}=1-\frac{(1-R^2)(N-1)}{N-k-1} \tag{9}$$

where k is the number of variables in the regression model.

Both R^2 and R^2_{adj} indicate the percentage of a dependent variable's variance that can be accounted for by the regression model's independent variables. However, $R^2_{adj}$is used as an evaluation of how

DOI: 10.54254/2755-2721/46/20241517

successful a regression model predicts responses for new observations. It will increase as more beneficial variables are incorporated into the model, and decrease reversely.

3.3. Performance comparison

Table 2. The result of different methods

Model	RMSLE	MAPE	RMSE	Adjusted R^2
SVM	0.243	0.432	1.56	0.821
Random Forest	0.212	0.243	1.44	0.892
XGBoost	0.193	0.210	1.12	0.912
Neural Network	0.172	0.167	0.98	0.922

From the table 2, Observations reveal that the Neural Network model, out of all the models, it functions the best., with the lowest values for MAPE, RMSE, and Adjusted $\boldsymbol{R^2}$, which are 0.167, 0.98, and 0.922, respectively. This indicates that the Neural Network model is capable of predicting suicide rates more accurately, demonstrating improved data fitting and lower RMSE and average absolute percentage error.

In contrast, the SVM model shows higher error values in terms of MAPE and RMSE (0.432 and 1.56, respectively), resulting in poorer accuracy. Although the model achieves relatively good results in RMSLE (0.243), there is still room for improvement in its performance in suicide rate prediction, possibly due to the model's insensitivity to samples with larger residuals.

Both the Random Forest and XGBoost models yield satisfactory results among all the models. They exhibit low values for RMSLE, MAPE, and RMSE, and perform well in terms of Adjusted $\boldsymbol{R^2}$. While the XGBoost model slightly lags behind the Neural Network model by 3% in terms of Adjusted $\boldsymbol{R^2}$, it still demonstrates good predictive performance.

In conclusion, the Neural Network and XGBoost models show good performance in predicting suicide rates, potentially because they are able to capture complex patterns and nonlinear relationships in the input data more effectively. On the other hand, the SVM model may perform relatively poorly due to its insensitivity to samples with larger residuals. Therefore, when selecting a suitable model for predicting suicide rates, it is crucial to consider accuracy, robustness, and the ability to fit the data's features.

4. Conclusion

In summary, this paper incorporates both machine learning and deep learning techniques to identify an appropriate approach for predicting suicide rates. Based on the analysis of different algorithms for predicting suicide rates, the following conclusions can be drawn:

The Neural Network model demonstrates the best performance in predicting suicide rates, with lower error values and higher fitting capabilities. With the continuous development of DL techniques, the NN model has shown strong capabilities in many prediction tasks and is expected to continue playing an important role in suicide rates prediction.

Both XGBoost and Random Forest, as ensemble learning methods, have achieved satisfactory results. They can effectively handle complex data relationships and serve as viable models for predicting suicide rates. Random Forest and XGBoost models have shown excellent predictive accuracy and robustness, although slightly lower performance of R2 compared to the Neural Network model.

In contrast, the SVM model performs relatively poorly in predicting suicide rates. Although the model achieves good results on certain metrics, its performance is subpar when dealing with samples with larger residuals, resulting in lower overall predictive accuracy compared to other models.

Looking ahead, research and practice in suicide prediction continue to advance. As big data and machine learning technologies have advanced, there is potential to build more accurate, reliable, and personalized suicide prediction models. Additionally, integrating data from other domains such as social media, biological markers, and genomics can enhance the accuracy of prediction models. Furthermore,

DOI: 10.54254/2755-2721/46/20241517

leveraging techniques such as natural language processing and sentiment analysis can further explore potential correlations and mechanisms of suicide rates, providing a comprehensive understanding for early prediction and intervention.

In conclusion, by applying different algorithms and models, valuable insights can be provided for suicide rates prediction, improving diagnostic accuracy and intervention effectiveness, and supporting future treatment research and personalized healthcare. However, continuous efforts are needed to improve and deepen prediction models to ensure their feasibility and applicability in practice.

References

[1] Noble W S. What is a support vector machine? Nature biotechnology, 2006, 24(12): 1565-1567.
[2] Xue H, Yang Q, Chen S. SVM: Support vector machines. The top ten algorithms in data mining, 2009, 6(3): 37-60.
[3] Chen T, Guestrin C. Xgboost: A scalable tree boosting system, Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016: 785-794.
[4] Chen T, He T, Benesty M, et al. Xgboost: extreme gradient boosting. R 0.4-2, 2015, 1(4): 1-4.
[5] Biau G, Scornet E. A random forest guided tour. Test, 2016, 25: 197-227.
[6] Speiser J L, Miller M E, Tooze J, et al. A comparison of random forest variable selection methods for classification prediction modelling. ESA, 2019, 134: 93-101.
[7] Biau G. Analysis of a random forests model. JMLR, 2012, 13: 1063-1095.
[8] Kriegeskorte N, Golan T. Neural network models and deep learning. Current Biology, 2019, 29(7): R231-R236.
[9] Scarselli F, Gori M, Tsoi A C, et al. The graph neural network model. TNN, 2008, 20(1): 61-80.
[10] Goldberg Y. A primer on neural network models for natural language processing. JAR, 2016, 57: 345-420.